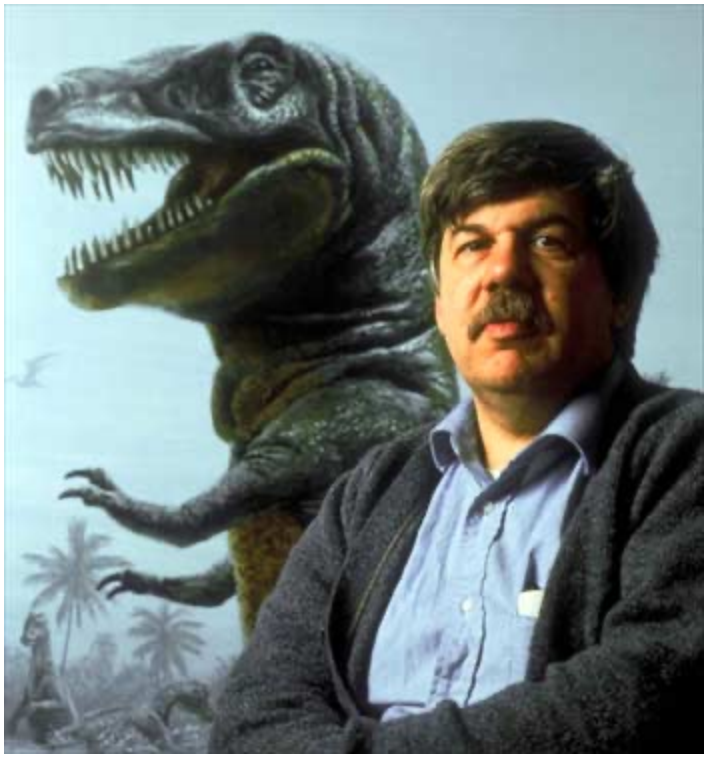
Molecular Epidemiology (First practical book 1993)
P. Eline Slagboom, biologist, Prof. of Molepi



Stephen Jay Gould, Evolutionary Biologist
1982; 41 years , abdominal mesothelioma
Prognosis: median survival is 8 months

I met him in 1993, the year of my PhD.

He died 20 years after diagnosis.
What happened.
Gould: “The Median is’nt the Message”

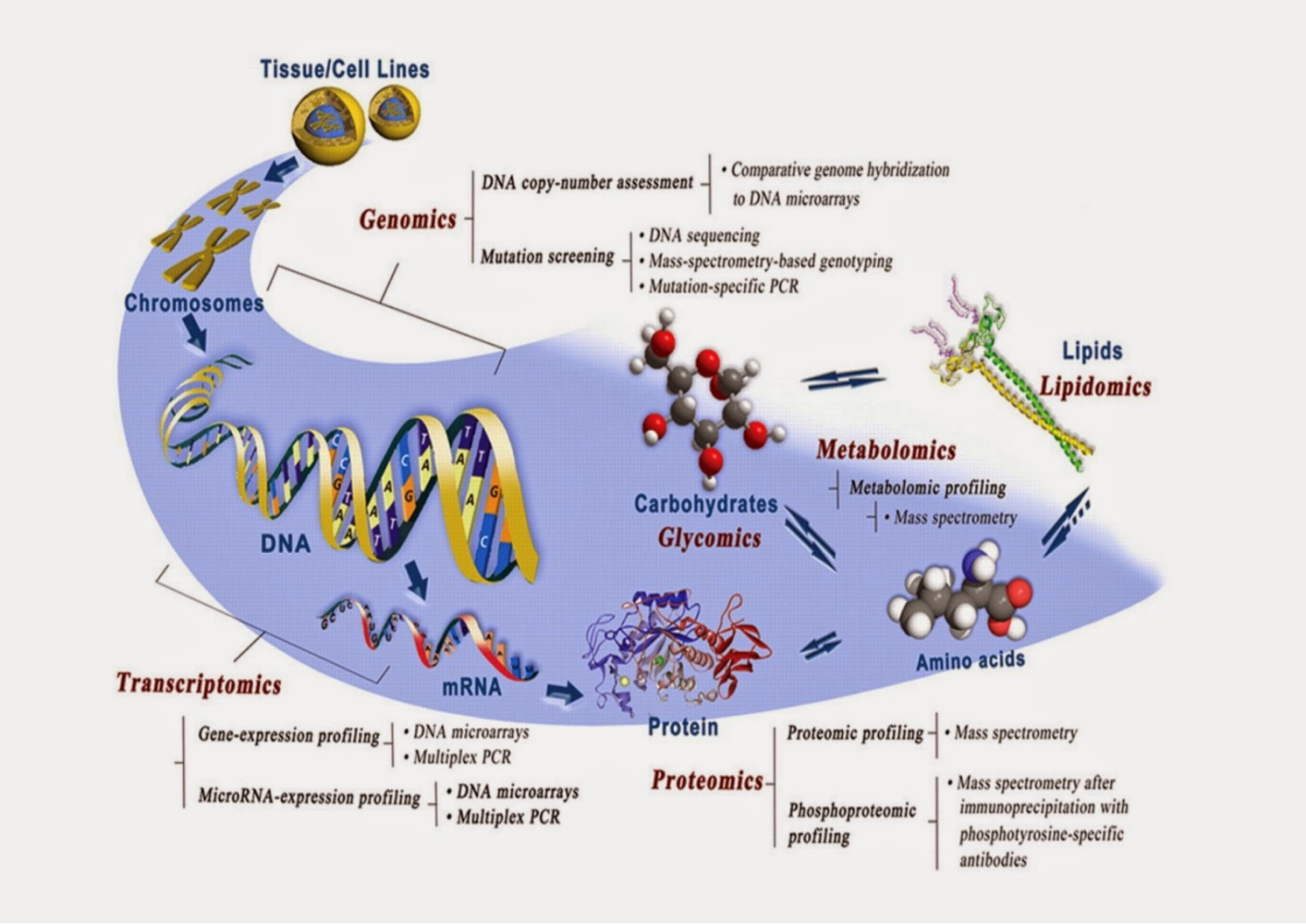
Molecular Epidemiology

- Introduced by Kilbourne (1973), infectious diseases; Schulte and Perera (1993 Principles and Practices)
- Integrates Epidemiology, Medical Sciences and Molecular Biology
- Studies the influence on health of environmental and genetic risk factors measured by (holistic) molecular signatures
- Contributes to
  prediction/prognosis
  monitoring exposure, response to interventions
  etiological understanding (disease mechanisms)

Since 1990: Revolution in Molecular Technology

high throughput recording exposure, consequence (biological change), pathway analysis.

Holistic: Targeted and Untargeted (hypothesis-driven and -free)

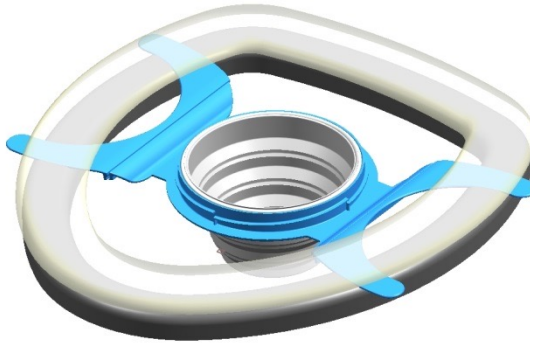


On Biological materials, molecular data and phenotypes (outcomes in studies)

What do we collect and what are study designs in which we collect data ?

Biomaterials: biopsies at operations; visits in studies; by mail

SOURCE	INPUT	DNA YIELD
BLOOD	200 µl	4-12 µg
BUFFYCOAT	200 µl	25-50 µg
MOUTH SWABS	1	3-10 µg
SALIVA	2 ML	110 µg
BONE	55-70 mg	5.5 µg
CARTILAGE	50-100 mg	2.4 µg



-80°C Biobank

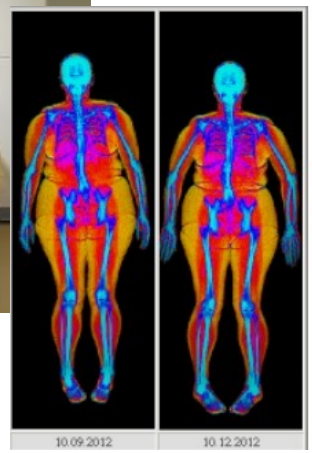
Biomateral of thousands of people: BIOBANK

Celmateriaal in de vloeibare stikstof opslagtank

-20°C Biobank

Robotics at Molecular Epidemiology

Phenotypes in patients and populations (biobanking)



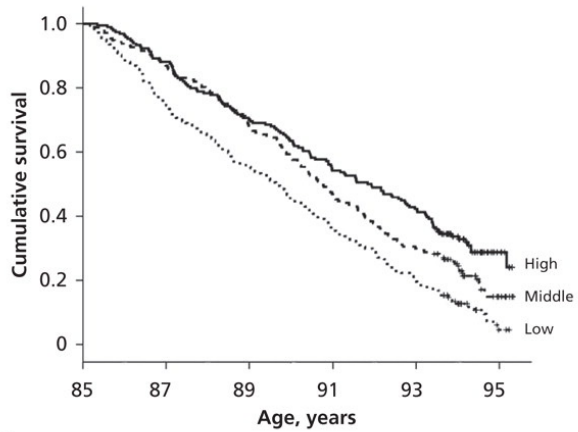
- Lifestyle, Demography
- Morbidity, Mortality
- Physiology

Functional

- Handgrip strength
- Cognitive functioning (memory, attention, speed, MMSE)
- Short Physical Performance Battery Protocol (gait, balance, chair rise)
- Questionnaires (sleep, quality of life, mood, depression, MMSE, 24 hrs food recall)

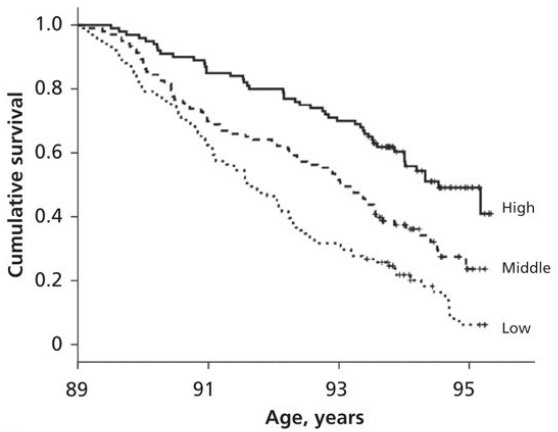
- Magnetische Resonantie Imaging (MRI)
- DEXA scans
- Wearable data
- Fluorenscence of the skin

A: Age 85 years



Tertile, *n*:						
High	194	171	136	105	82	61
Middle	177	154	119	82	53	34
Low	184	137	102	65	36	17

B: Age 89 years



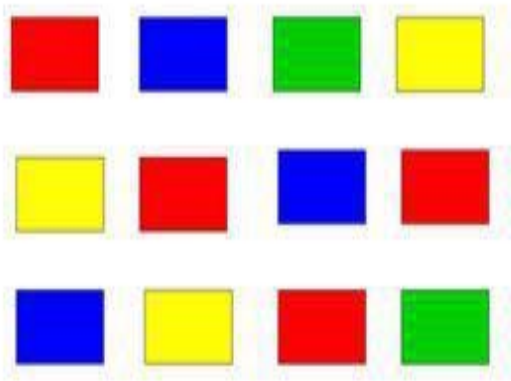
Tertile, *n*:				
High	100	85	70	54
Middle	103	72	53	33
Low	101	62	32	15

Grip strength in the Leiden 85 plus study

Test Cognition : STROOP

Stroop Taak: “Noem zo snel mogelijk de kleur”

Fase 1



Fase 2

Rood Blauw Groen Geel

Geel Rood Blauw Rood

Blauw Geel Rood Groen

Fase 3

Rood Blauw Groen Geel

Geel Rood Blauw Rood

Blauw Geel Rood Groen

Complex traits: effect of multiple genes and multiple environmental factors

What are study designs in which we collect data ?

Exposure (determinant) and Outcome; epidemiological study designs

http://cisncancer.org/research/images/epi_study_types-12-w.jpg

Study designs that now include MolEpi

Observational studies (cohorts, patients, twins)

- **Cross sectional (Case-Control)**
- **Prospective (longitudinally repeated measures rare)**

Experimental studies in humans (RCT)

- **Response to treatment**
- **Recording exposure**
  (to medication, food; in blood, urine, faecal samples)
- **in depth biological studies (i.e. cell biology)**

Complex traits: effect of multiple genes and multiple environmental factors

Is the trait or an element in it heritable ?

How to calculate the genetic and environmental component of any phenotype ?

- Comparison of phenotype in MZ and DZ twins

  Heritability $h^2 = 2*(rMZ-rDZ)$

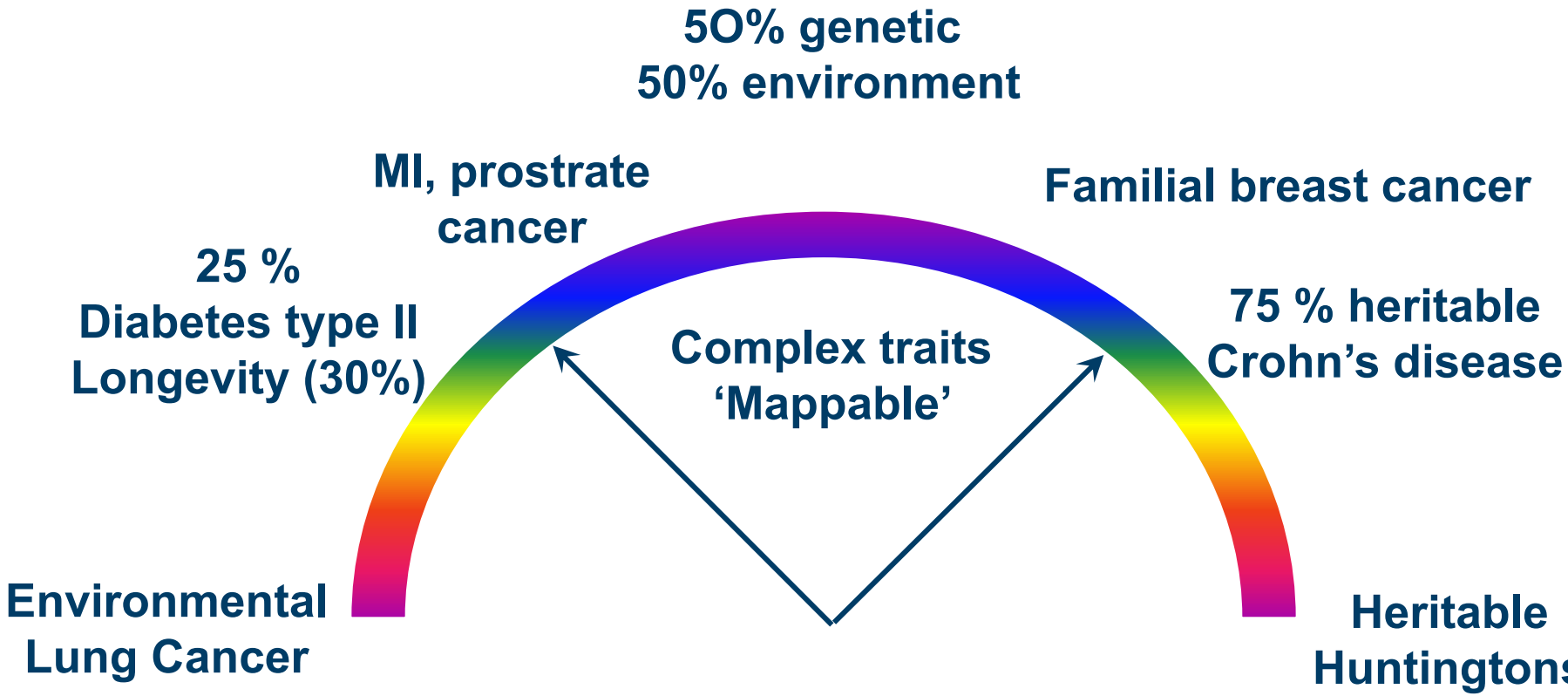
  Heritability of human longevity 33%

- Comparison of phenotype in family members

$$\lambda_{s=} \quad \frac{\text{risk for a sibling of an affected proband}}{\text{risk in the general population}}$$

cystic fibrosis $\lambda_{s=}500$ schizophrenia $\lambda_{s=}8.6$

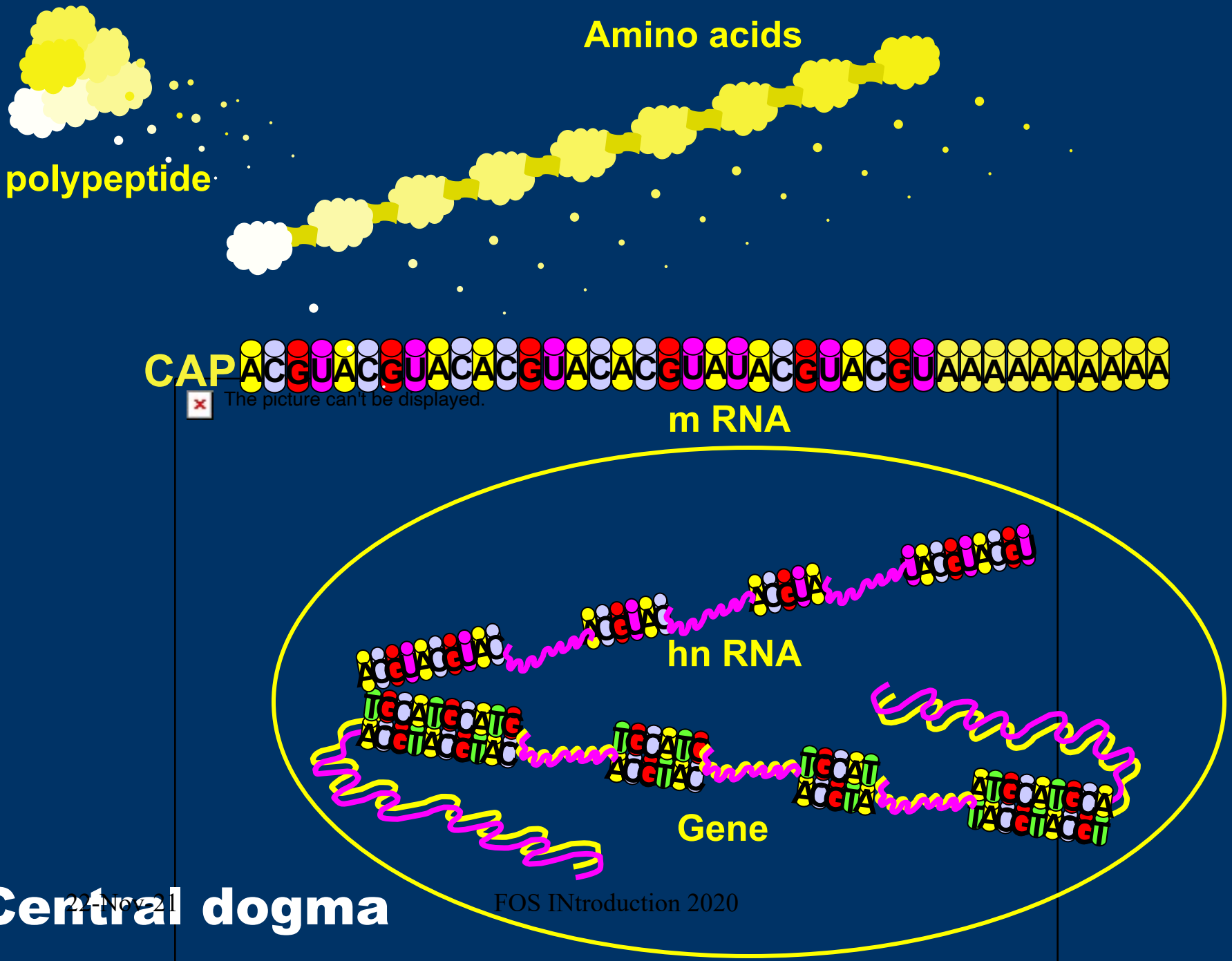
Heritability for common human disease



Mens: 22 chromosoom paren (autosomen) en sex chromosomen XX or XY

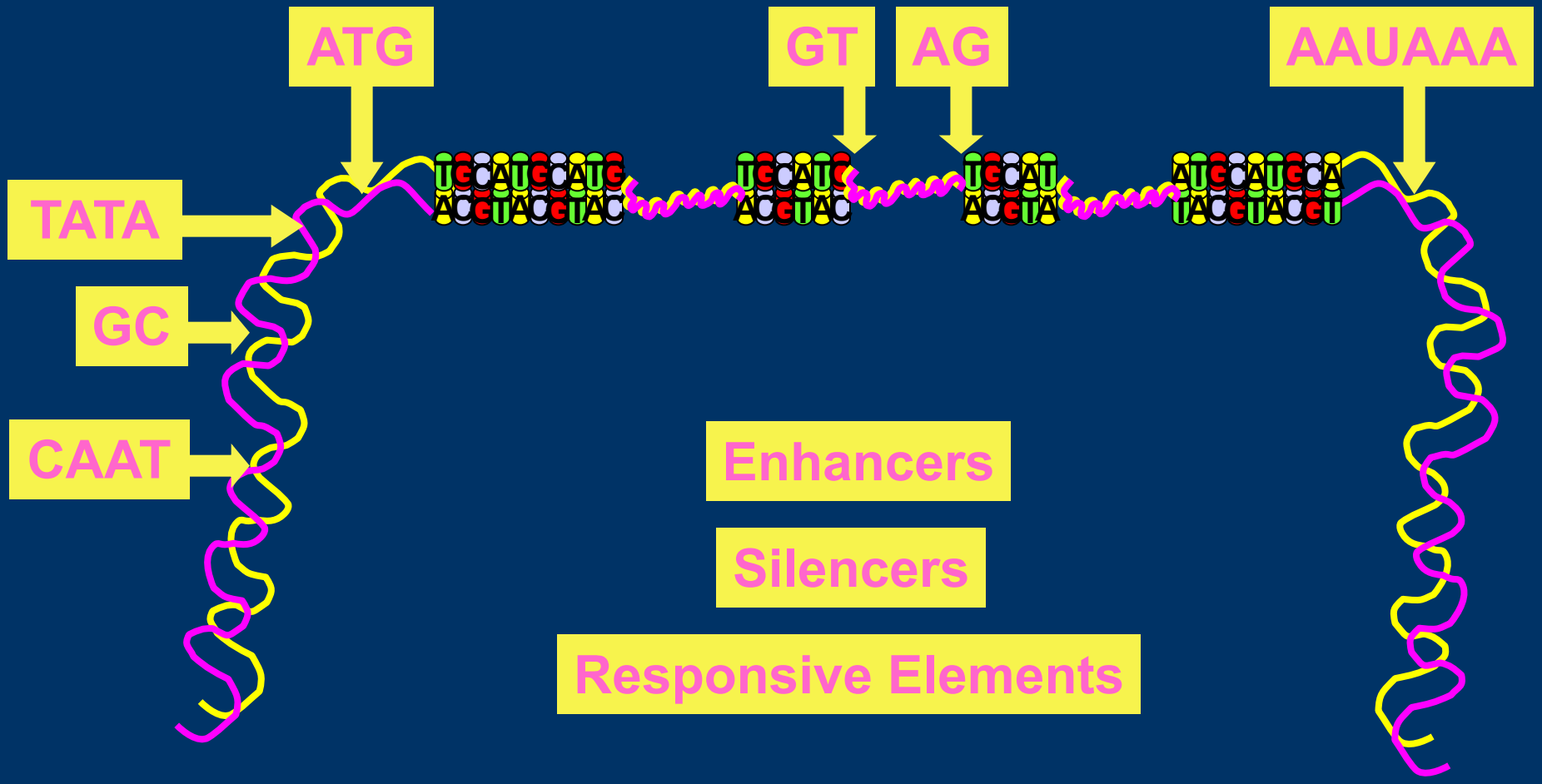
20.000 genen = 3% van al het DNA in uw cellen codeert voor een eiwit

What type of genetic variation may affect Complex or late life disease and ageing



Central dogma

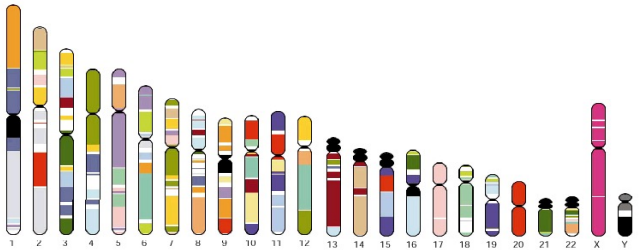
Signals for gene expression



Genetic variation with mild effect: not expected in exons

If there is a genetic component in a trait , how do we find the genes involved ?
Make study designs and collect biomaterial

Approaches to study how genetic variation contributes to disease (hypothesis driven and h-free, discovery science)



Genome wide association scan GWAS

Hypothesis generating
Genomic architecture (Explorative)
Linkage analysis (<5%)
Association analysis (>10%)

Candidate gene

'Hypothesis'-driven
Association analysis

DNA sequence

'Hypothesis'-driven
Next generation sequencing
whole genome/exome

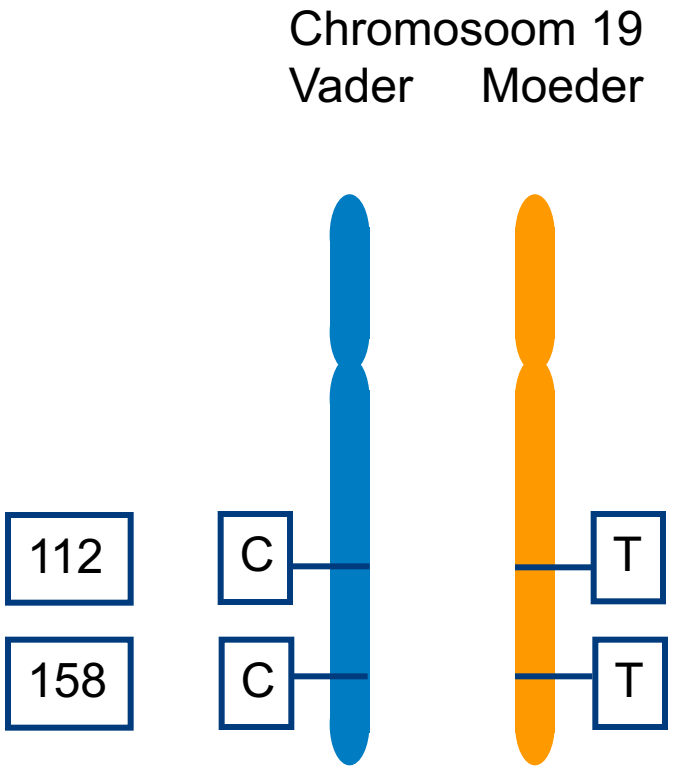
SNP = Single Nucleotide Polymorphism

Position 112 : allele C and T
Position 158 : allele C and T

Genotype 112 : CT

Haplotype 112-158 Father : C-C

Haplotype 112-158 Mother: T-T

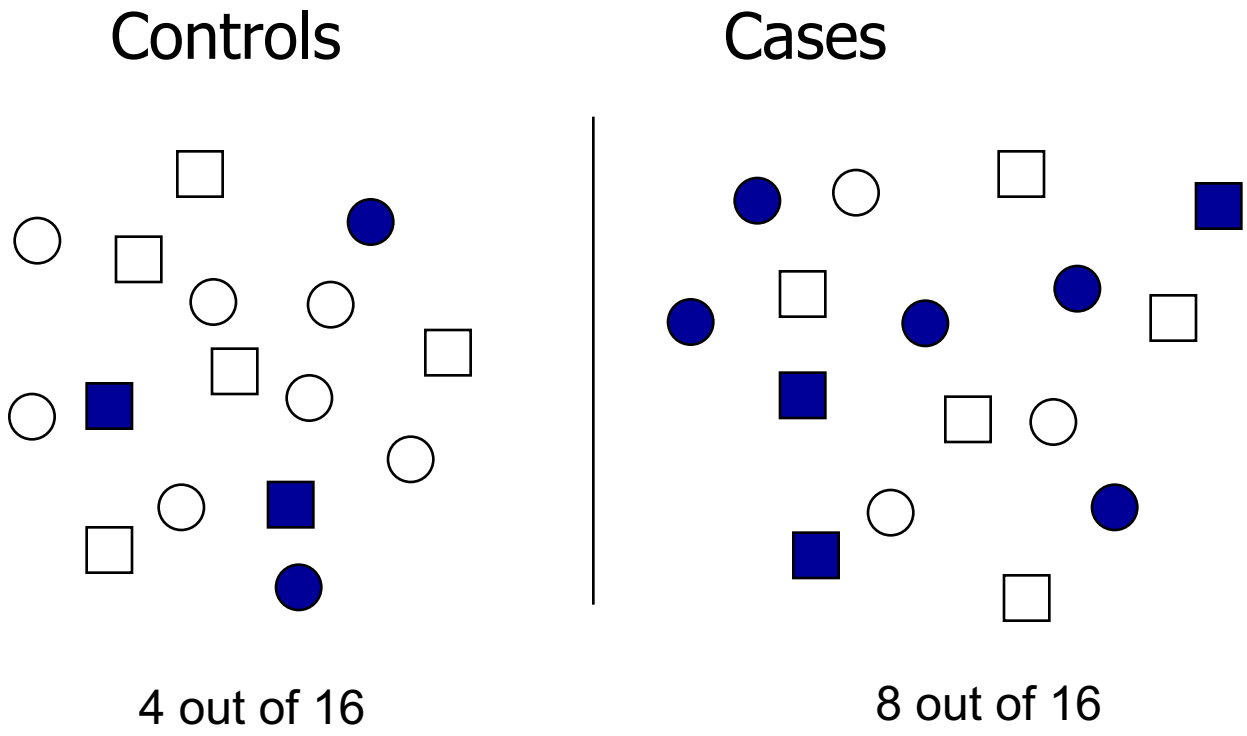


6 billion basepairs

150 million polymorphic positions In the human genome known

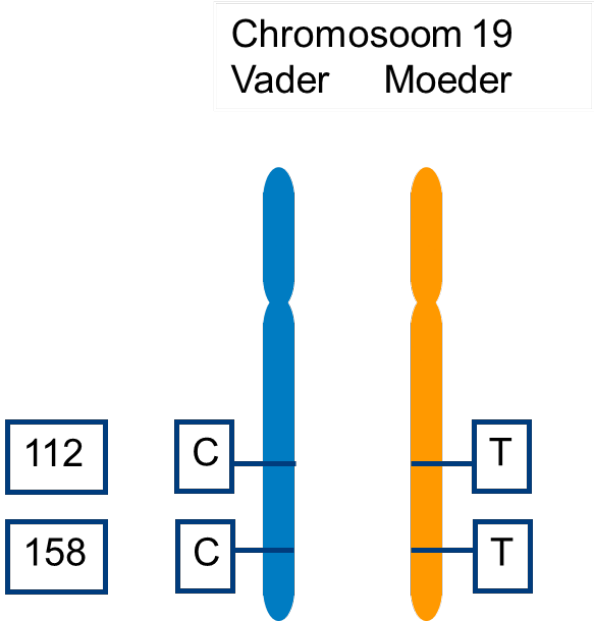
Genetic association study : Unrelated subjects

Genotype frequency in cases versus controls
Allele 2 is risk allele; count 12 genotypes



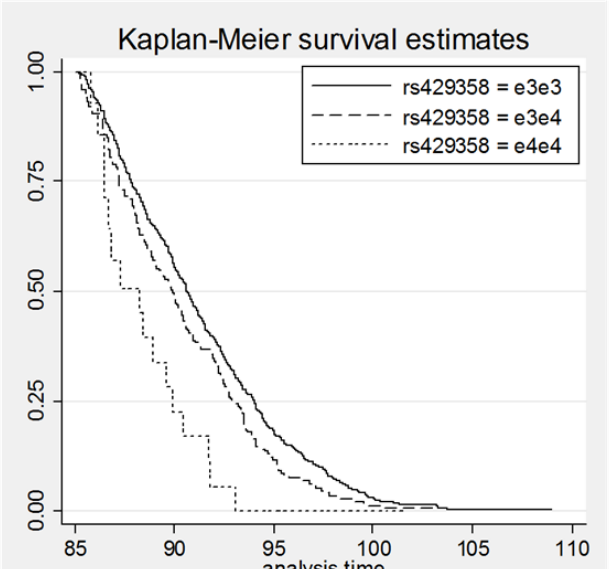
FUNCTIONAL VARIATION
APOEε2,3,4 locus

Haplotype
- APOEƐ2 : T-T
- APOEƐ3 : T-C
- APOEƐ4 : C-C



Ɛ2 : 8% is carrier ;
Associates with longevity

Ɛ4: 14% is carrier;
Associates with higher mortality risk and dementia
Homozygotes 15 times
Increased risk of dementia



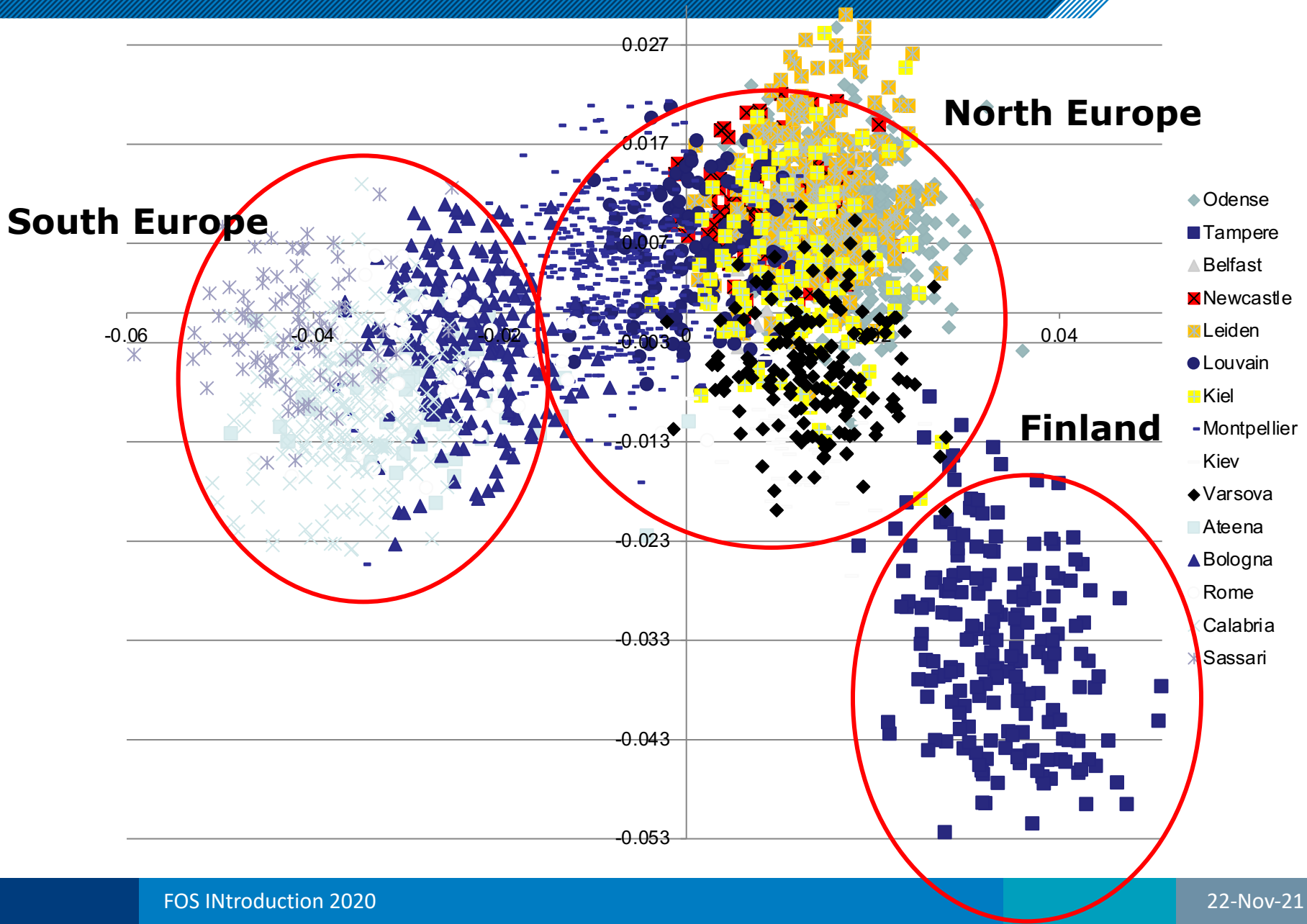
APOE in Leiden 85+ Study
15 years to follow up

Genome wide association study (GWAS) : search for genomic positions involved in the trait; then find the causal variant

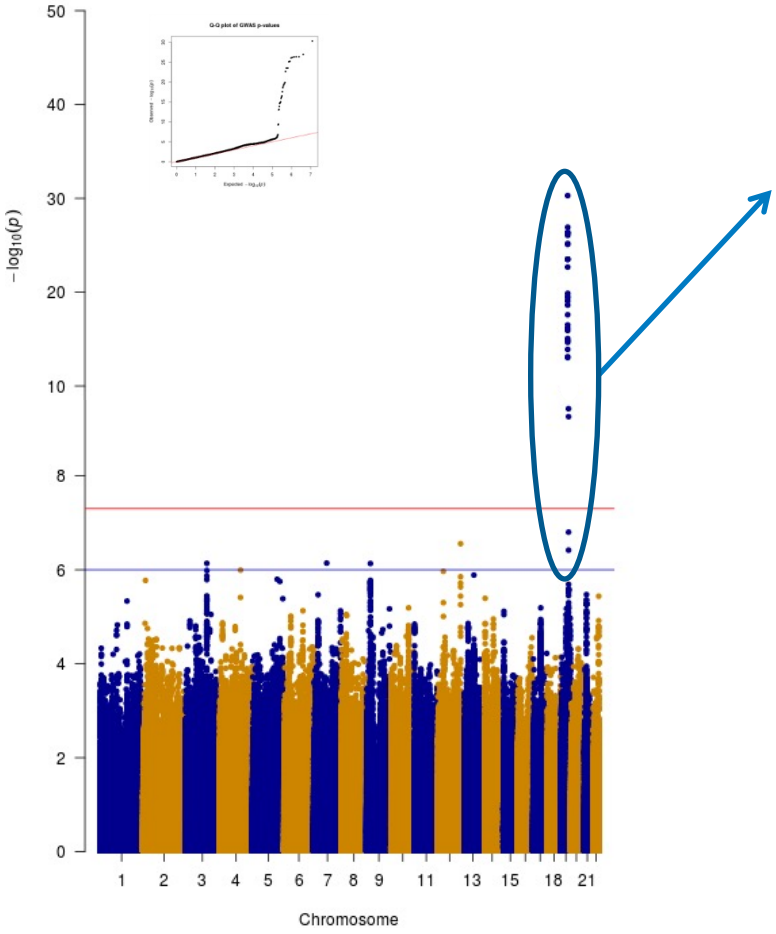
Added value of genome wide association studies (GWAS; scan million markers); also problems ?

Genome Wide Scan

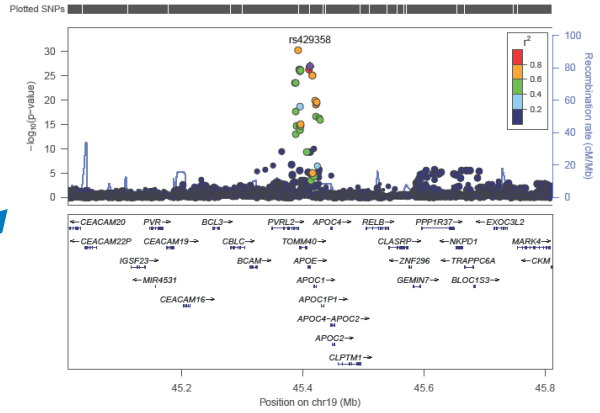
Genetic Origin: Stratification, a problem in genetic studies



Longevity : Finding the APOE locus in a GWAS



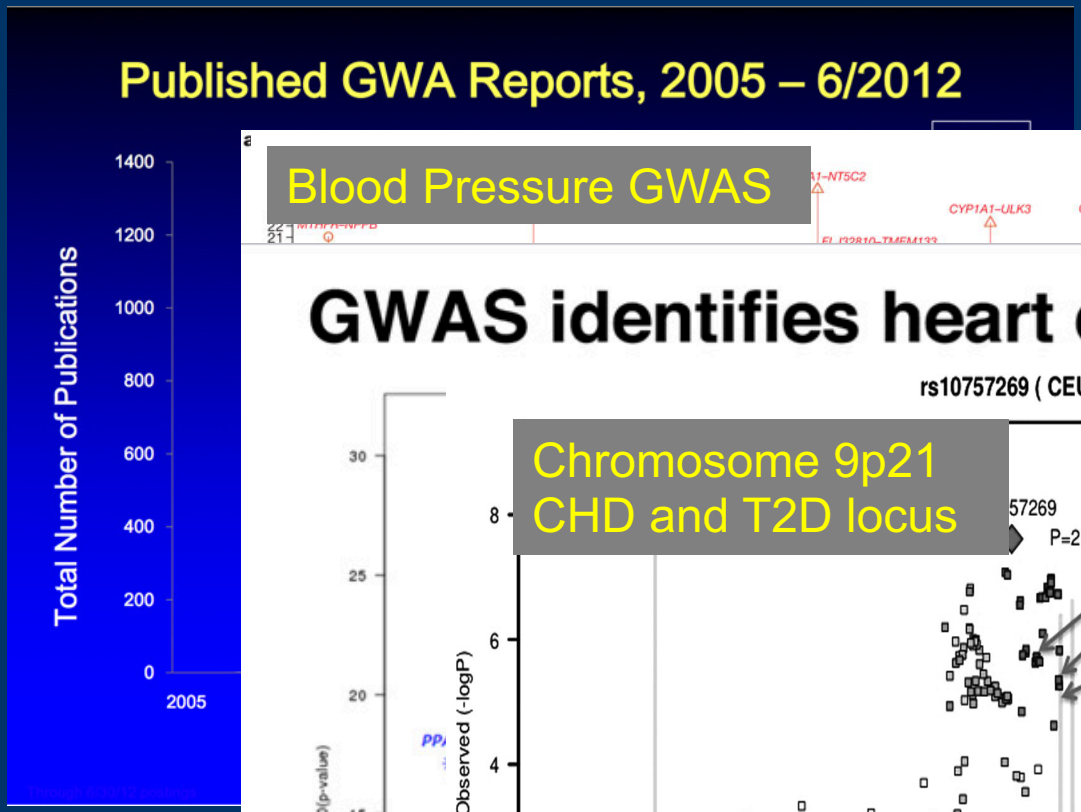
Manhattan plot



Interpretation of high dimensional data; Multiple testing

P 0.05 / no of tests (million) (Bonferroni correction

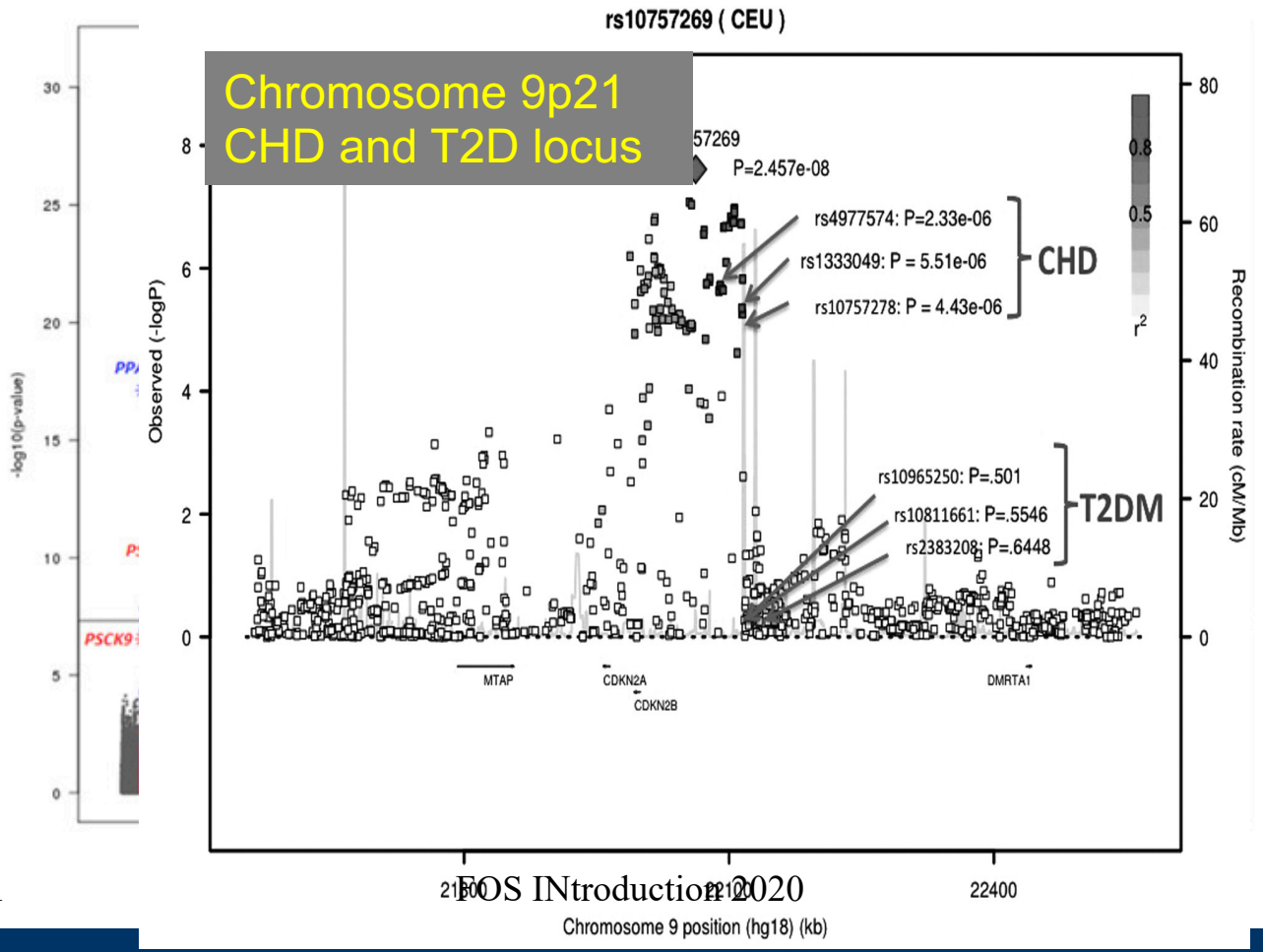
GWAS to find the location of disease susceptibility genes



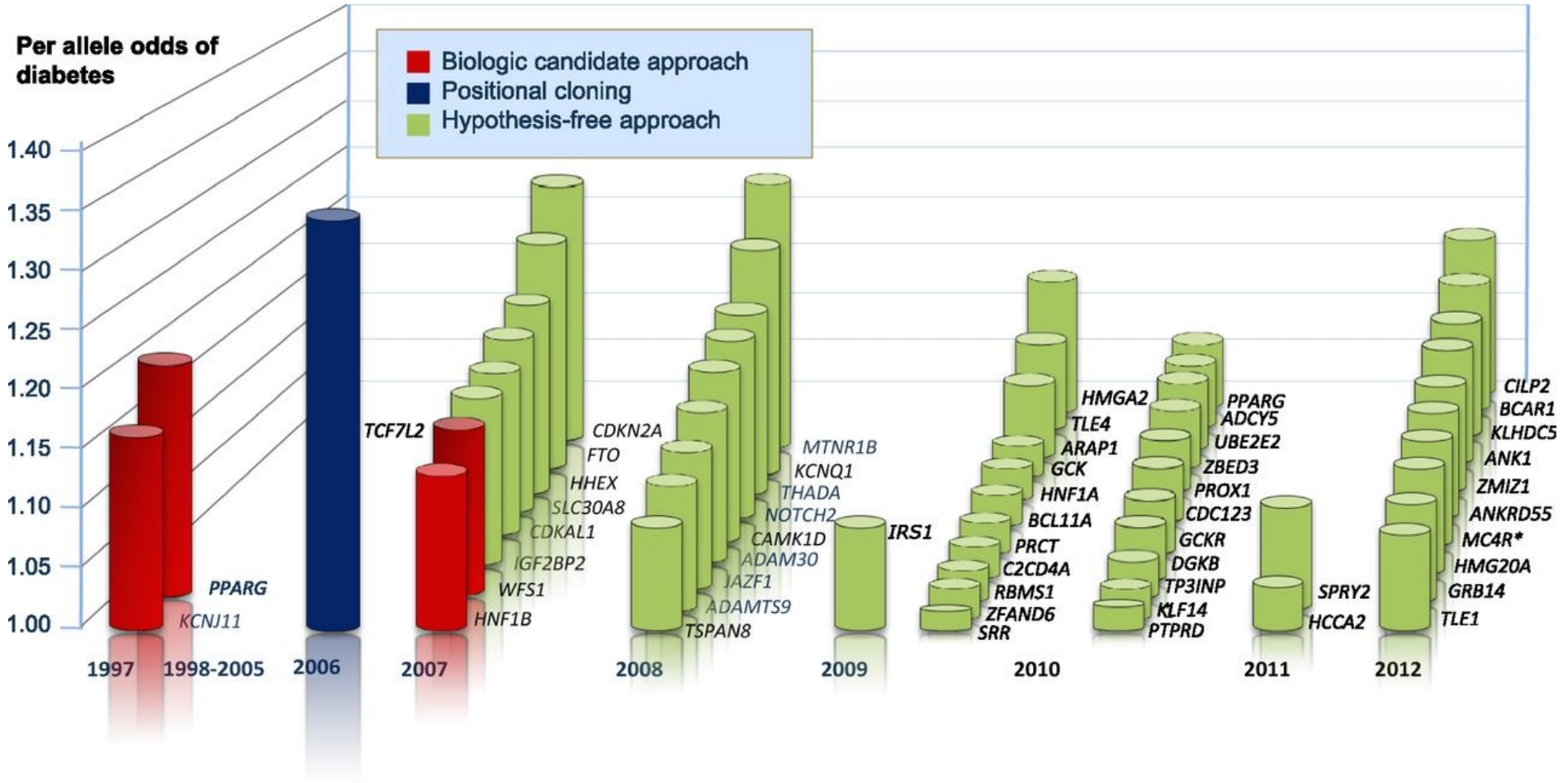
Blood Pressure GWAS

GWAS identifies heart disease factors

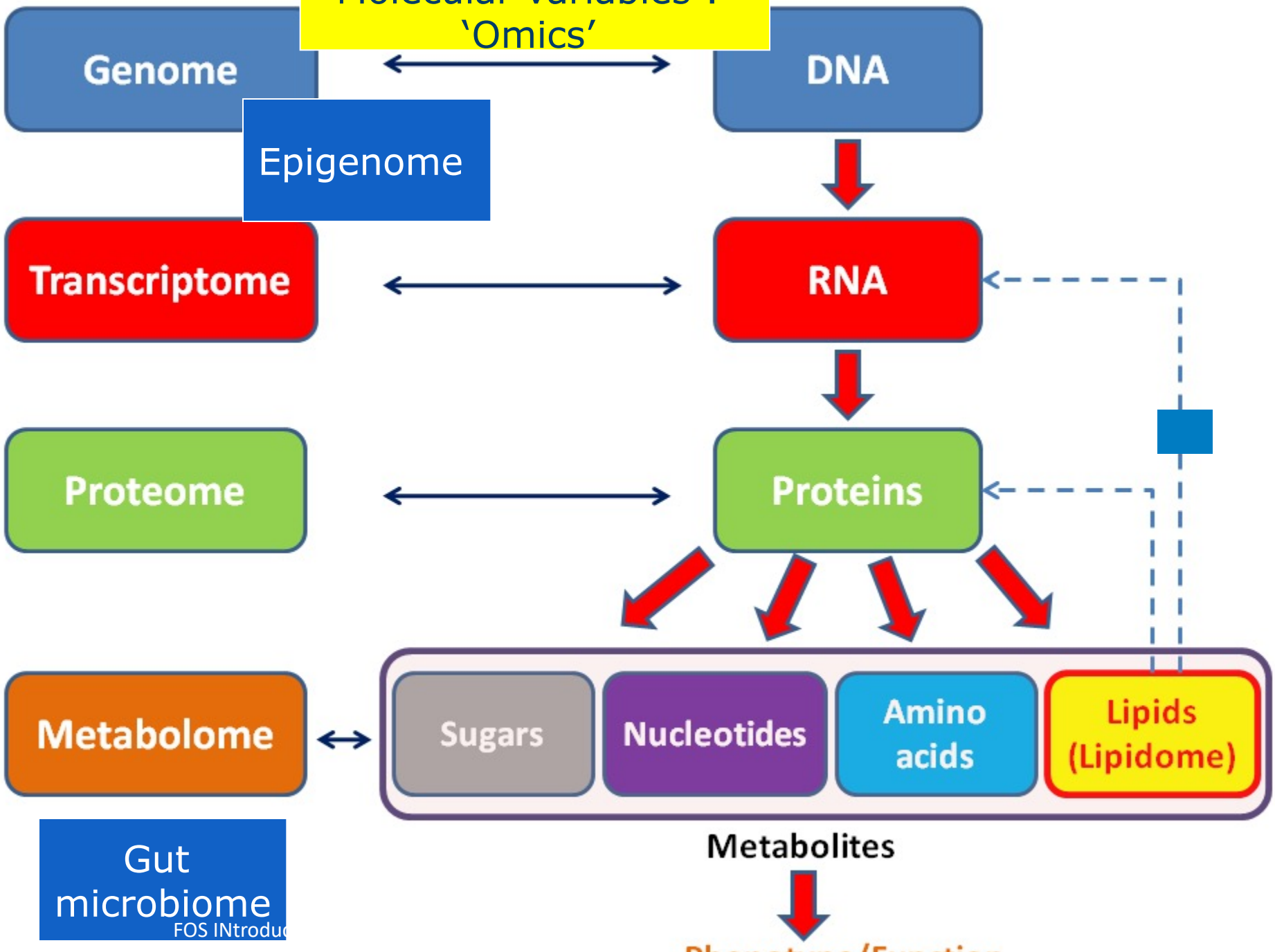
Chromosome 9p21
CHD and T2D locus



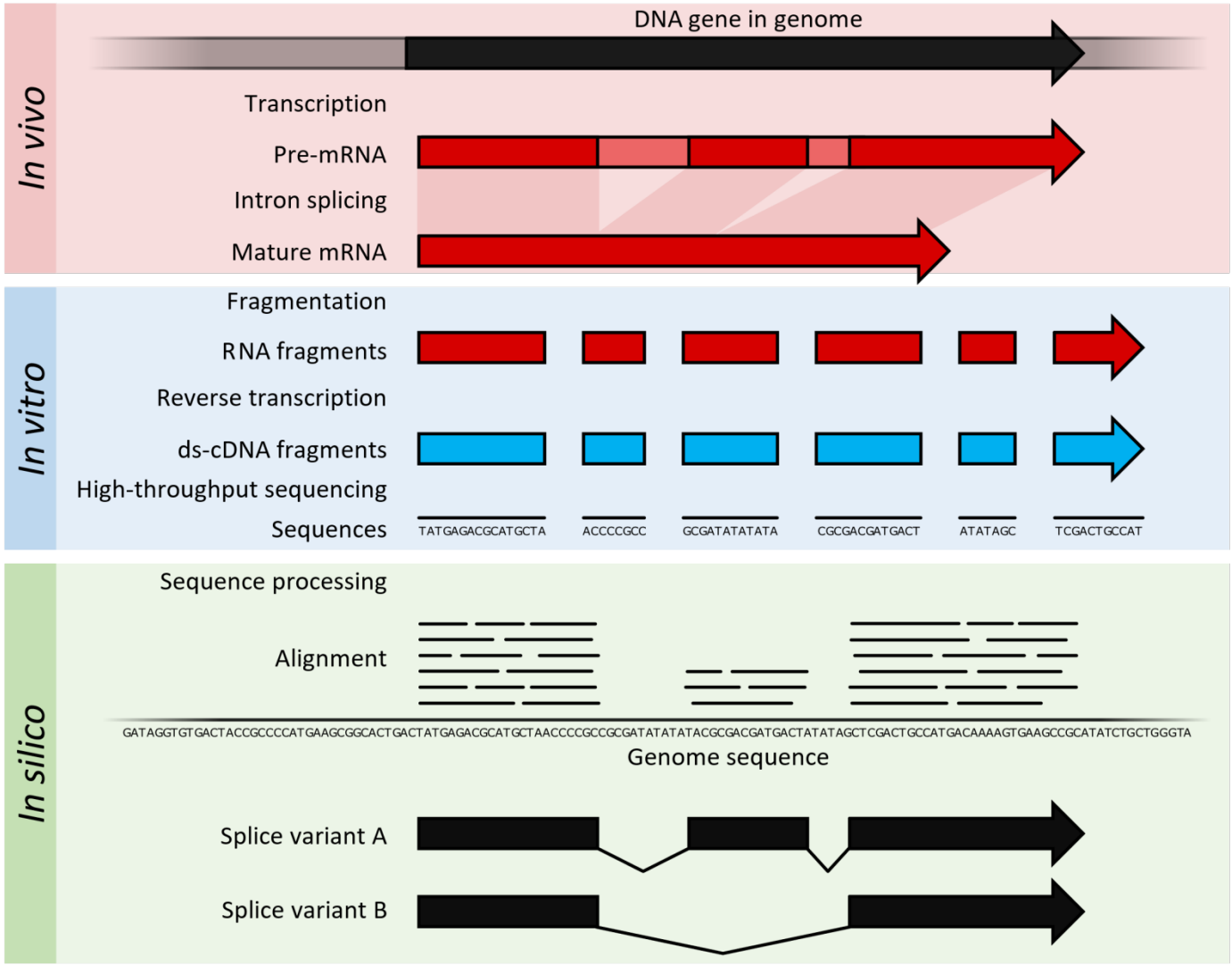
Type 2 Diabetes loci identified with different strategies
Calculate joint effect → Polygenetic Risk Score (PGRS or GRS):
20 % of variation in disease risk (only in European populations)



Molecular variables : 'Omics'



Gene expression.
Quantitative data (20.000 genes, different expression levels)



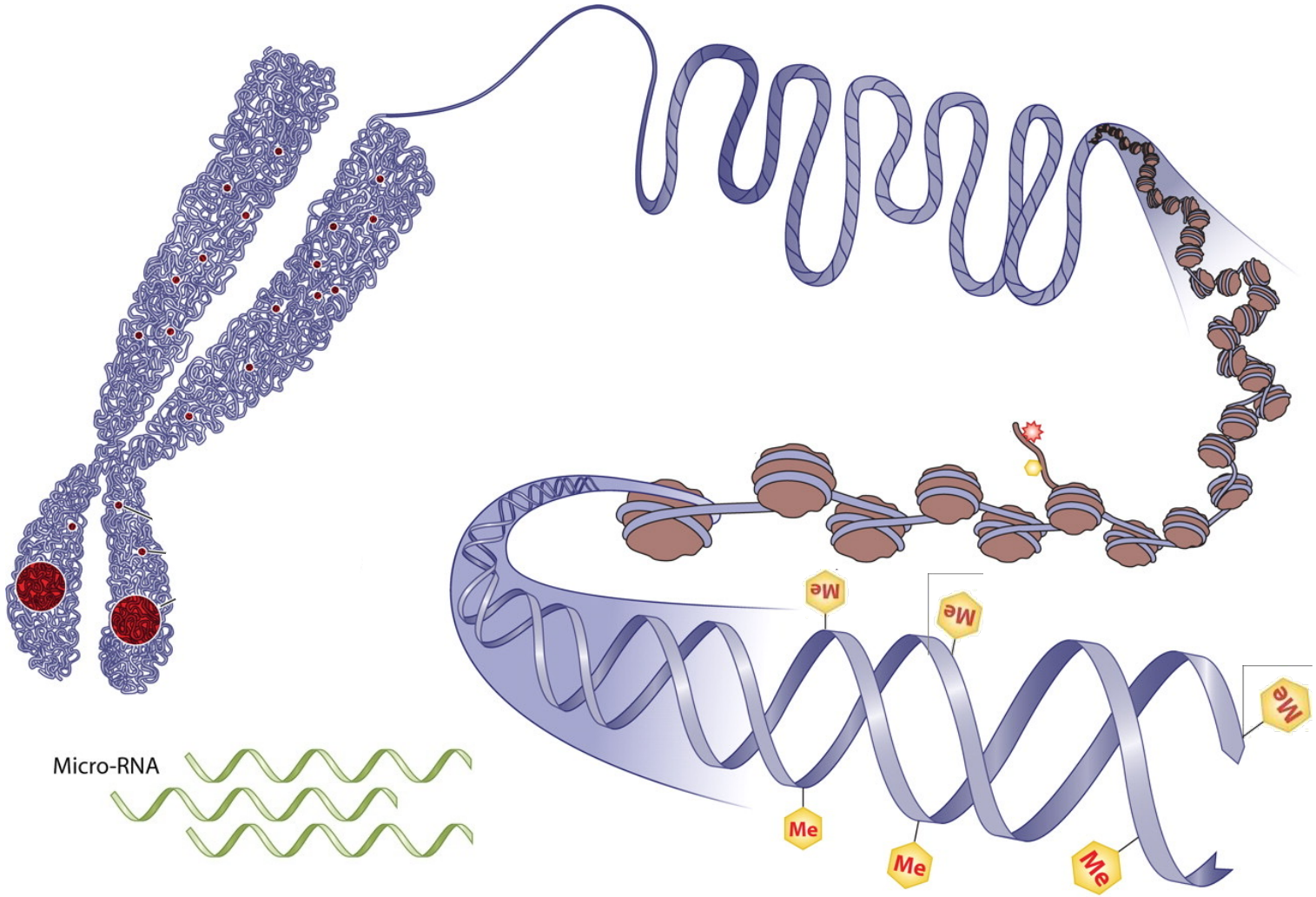
Gene expression variation (the transcriptome)

Expression of a single gene,
To 20.000 coding genes (coding for proteins)
(2% of the genome)

Non coding genes: 22 K
Transcripts : 98 K

Quantitative data

Genome records the environment : Epigenome



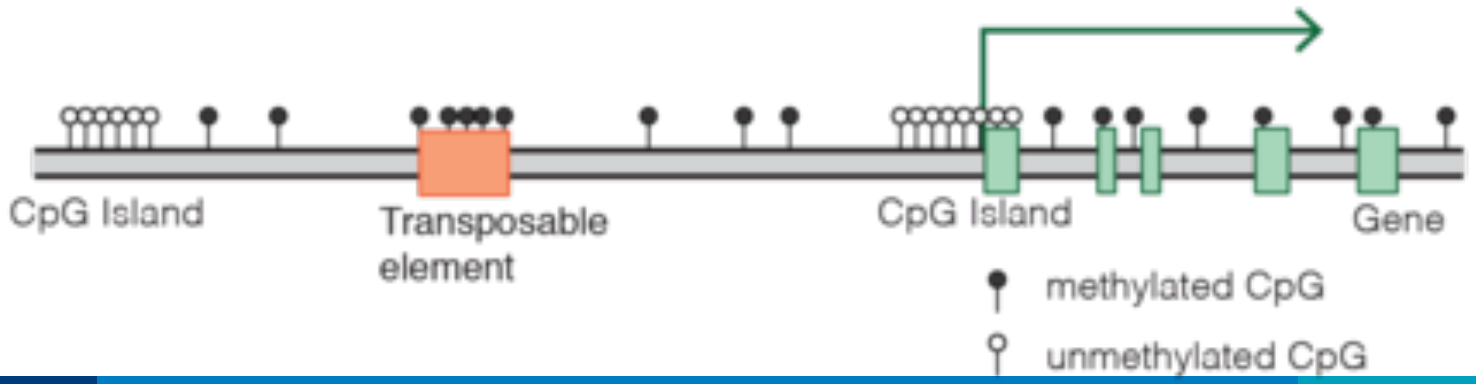
Epigenome

Most population studies into DNA methylation or non coding RNA

Smaller studies into chromatin configuration

Quantitative data
Genome wide DNA methylation assays:
27.000 to 850.000 CpG sites now

Typical mammalian DNA methylation landscape



BBMRI Biobanking consortium
Multi-level omics data set

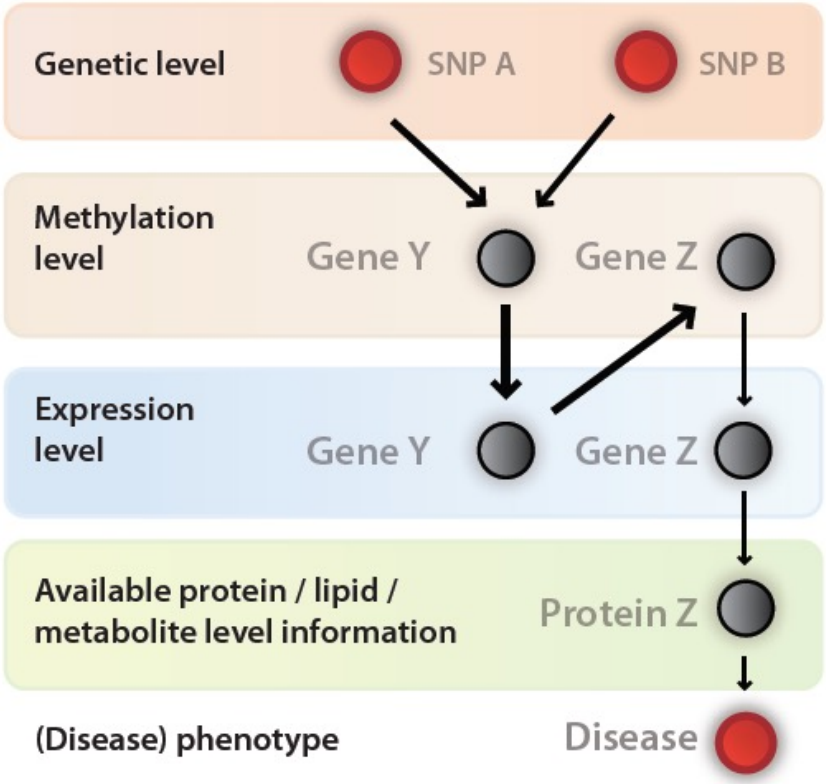
N=100,000 GWAS
N=750 Go.NlL

N=4,000

N=4,000

N=50,000

N>250,000



Go•NL
GENOMEoftheNETHERLANDS

BIOS consortium

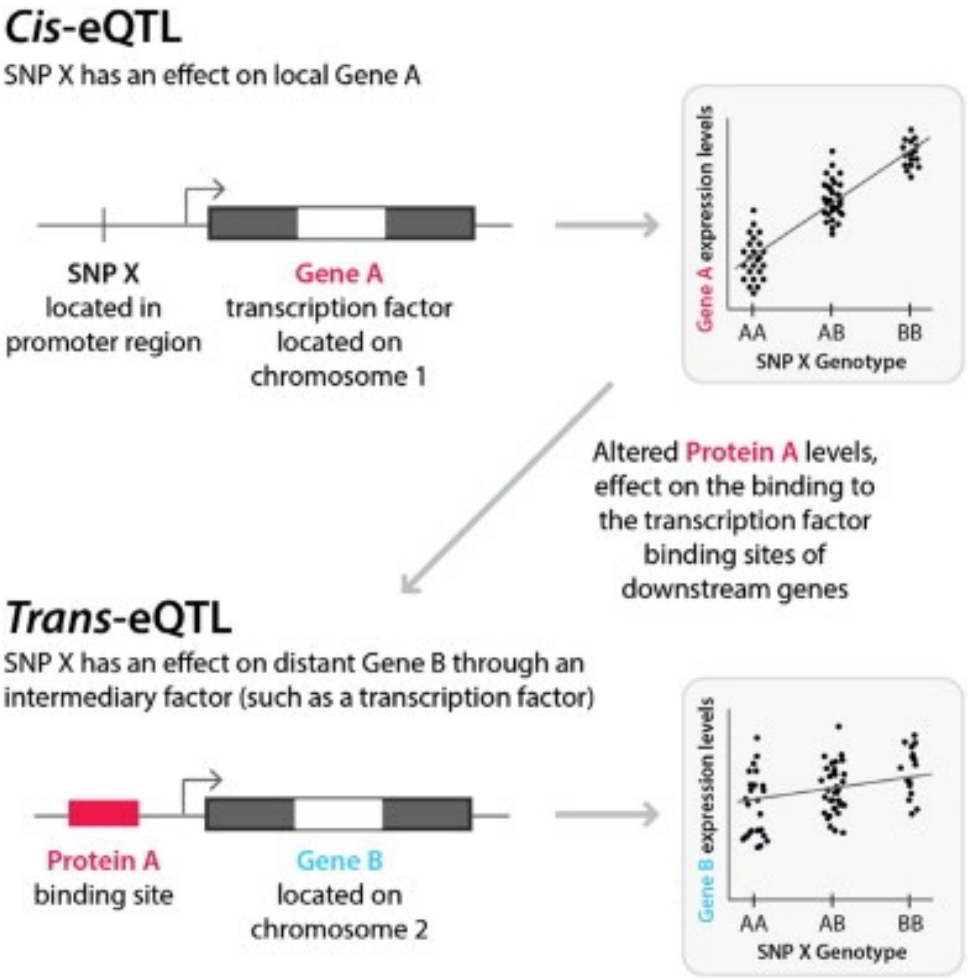
BIOS consortium

METABOLITE consortium

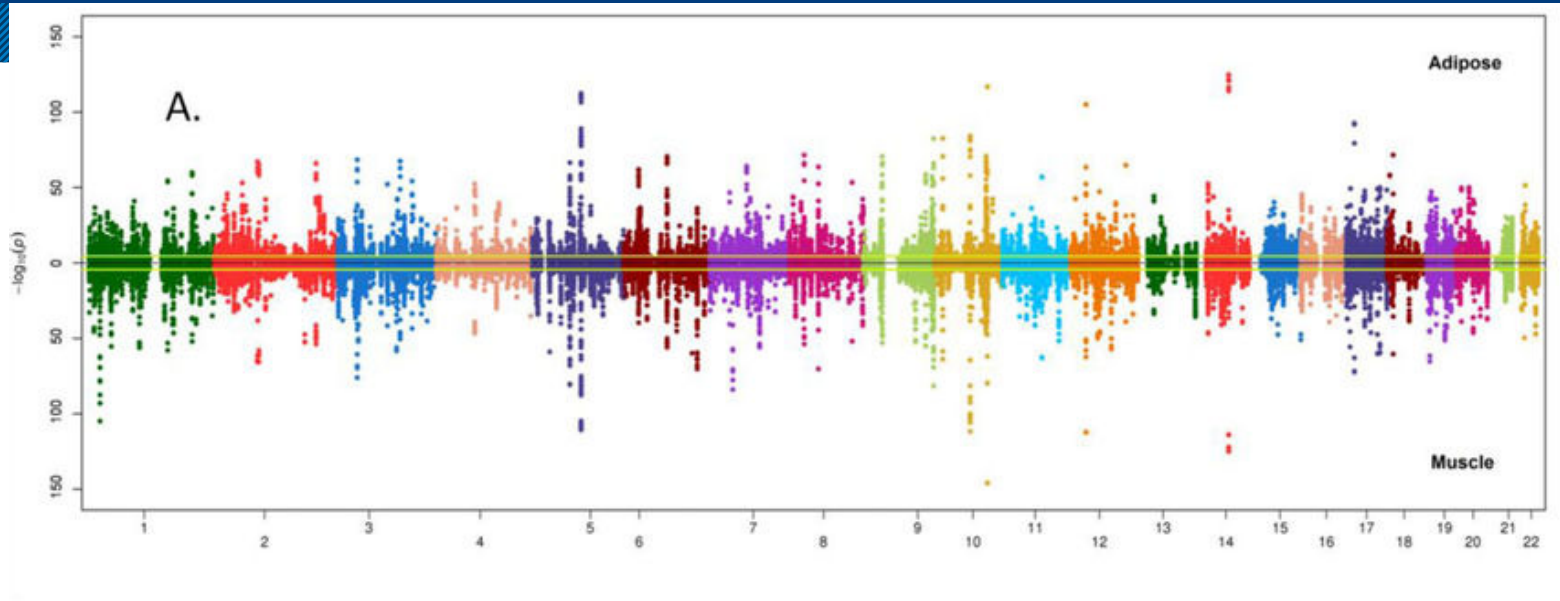
Central databases for the research community

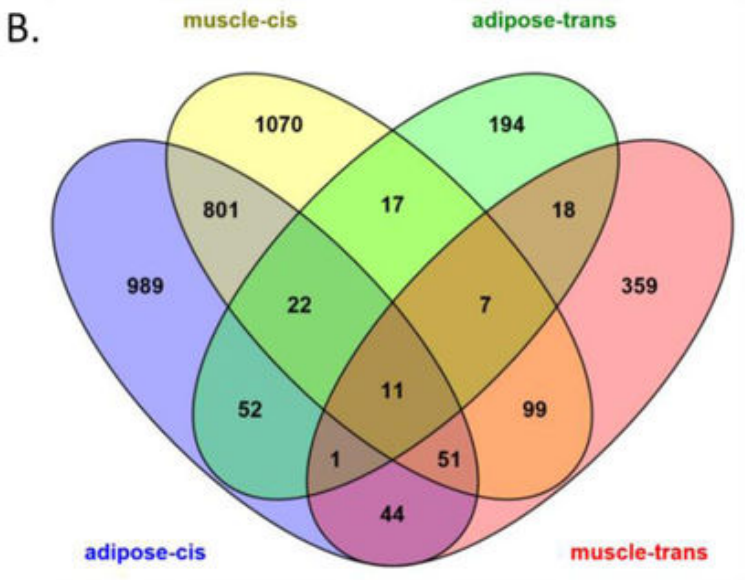
Genetic variation X quantitative variable (Quantitative Trait Loci (QTL) Quantitative data (20.000 genes, expression levels)

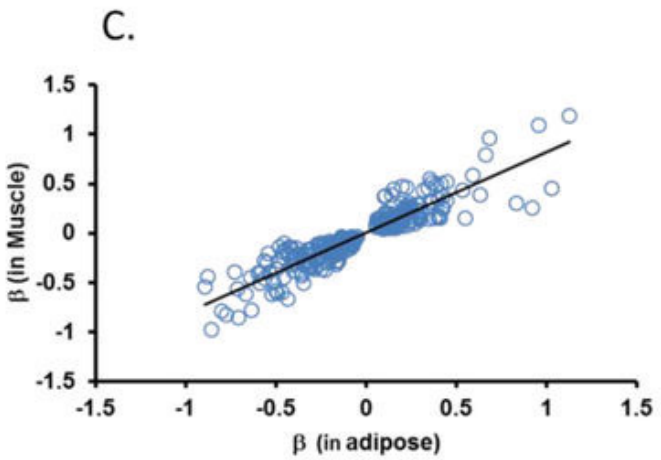
Expression QTLs (eQTL)



Scan the genome for SNPs/alleles influencing gene expression

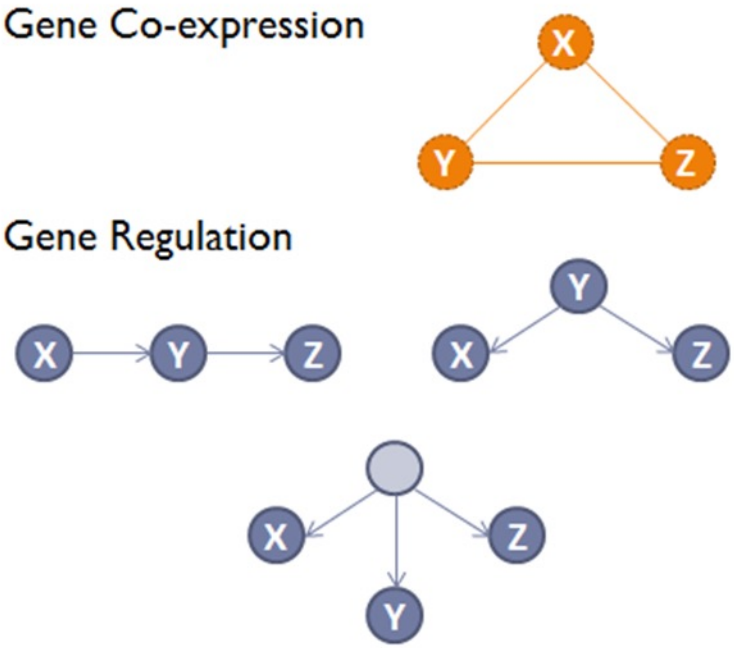




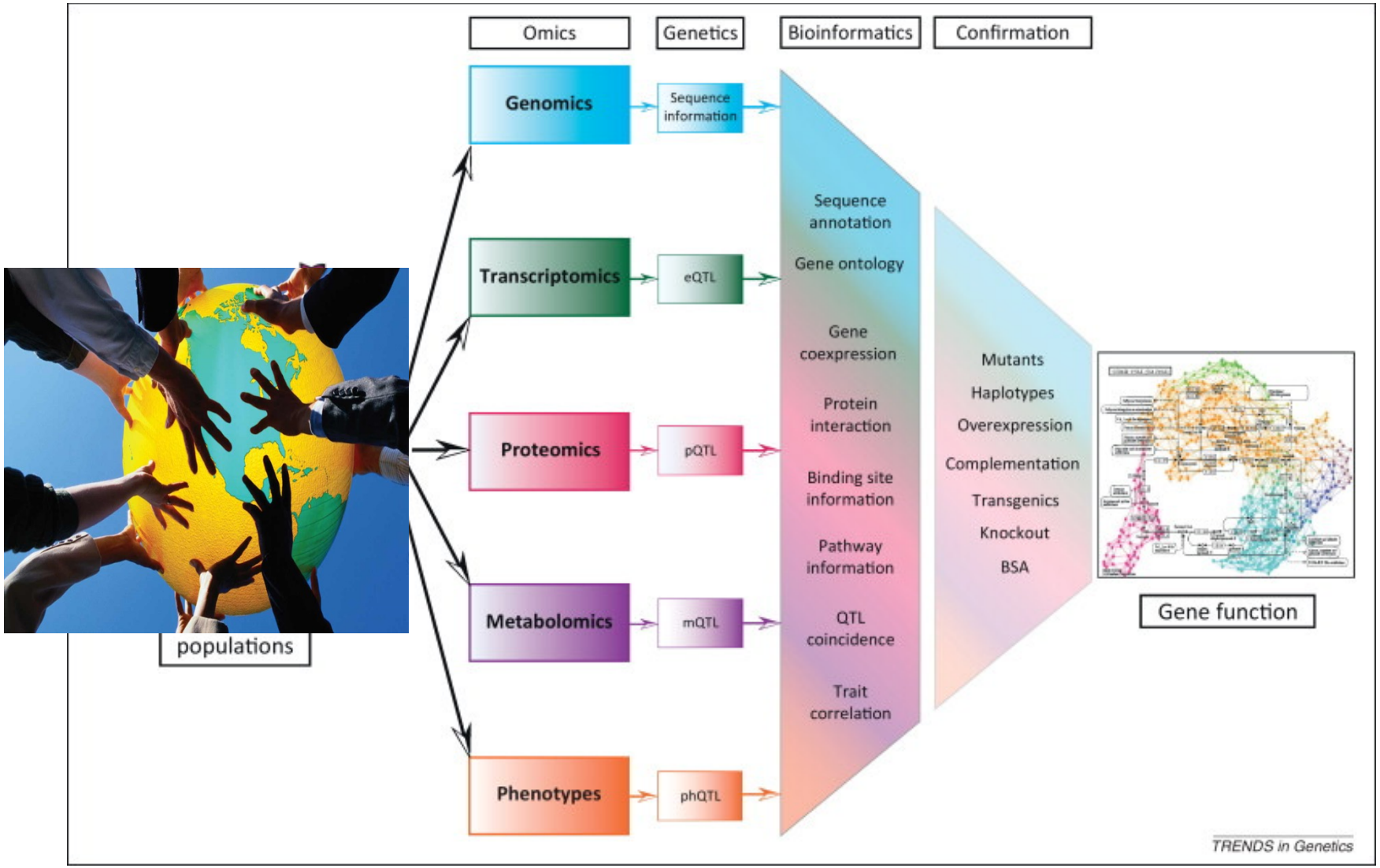


Expression QTLs (eQTL) in muscle and fat

Why is high dimensional molecular data complex ?

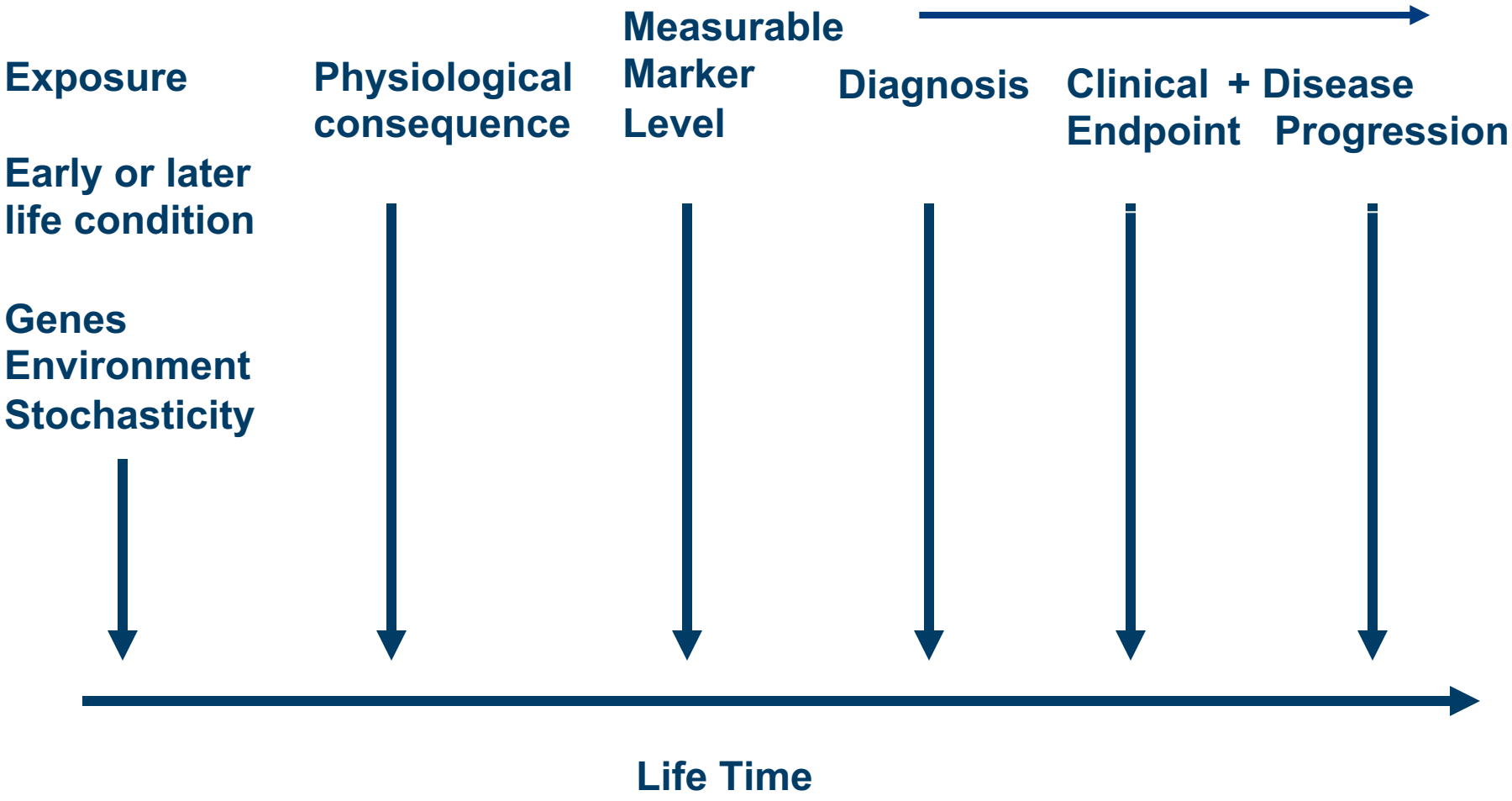


1. Multiple testing
2. Variables not independent
3. Correlation in the data
4. Distribution
5. Visualization of results (associations)


Omics
Genetics
Bioinformatics
Confirmation
Genomics
Transcriptomics
Proteomics
Metabolomics
Phenotypes
Sequence information
eQTL
pQTL
mQTL
phQTL
Sequence annotation
Gene ontology
Gene coexpression
Protein interaction
Binding site information
Pathway information
QTL coincidence
Trait correlation
Mutants
Haplotypes
Overexpression
Complementation
Transgenics
Knockout
BSA
populations
Gene function
TRENDS in Genetics

Sofar for etiological studies.
Molecular data can also help in recording exposure and prediction of an outcome.

Exposure Events in lifetime perspective



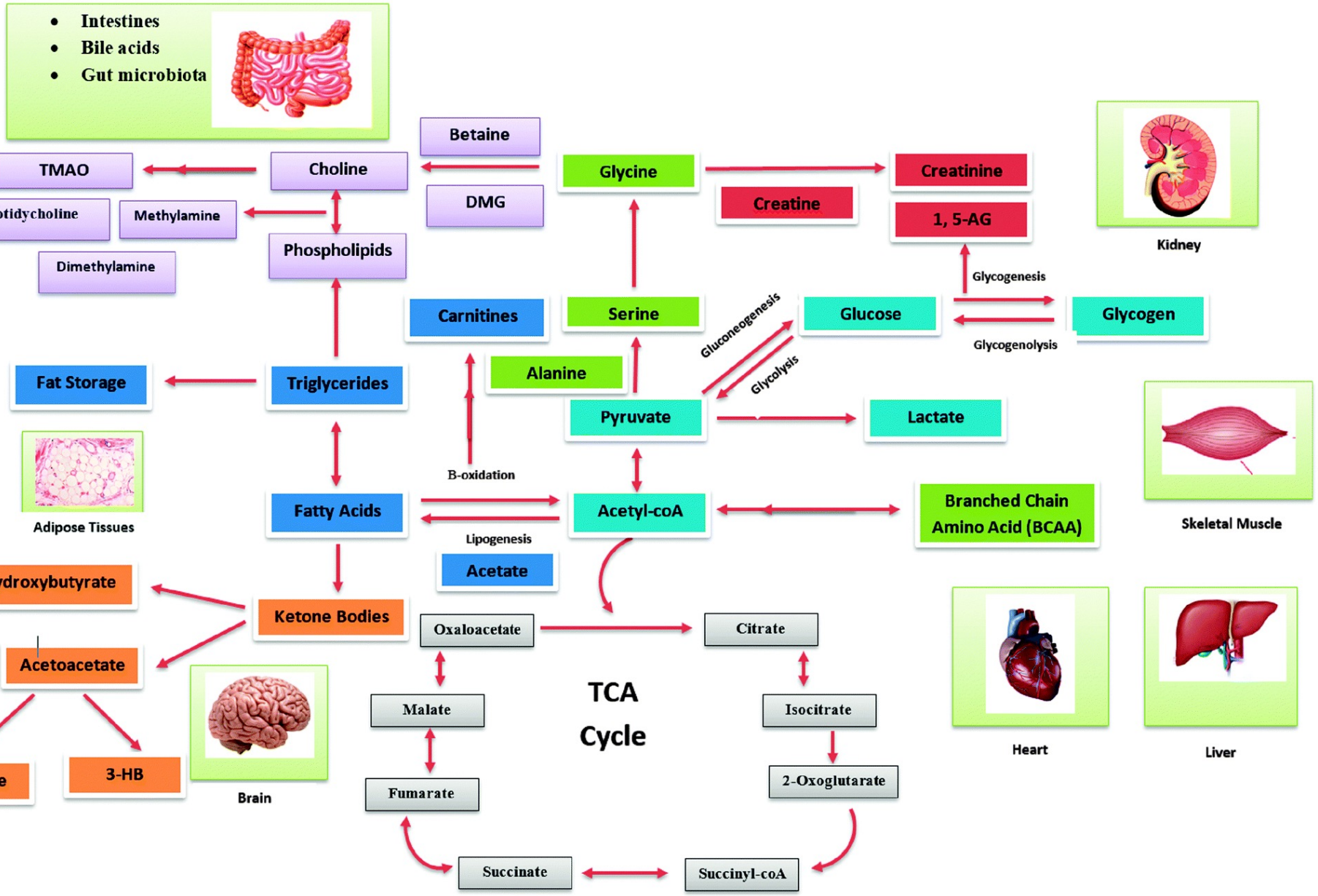
Biomarkers

- Relation of Exposure /determinant and outcome

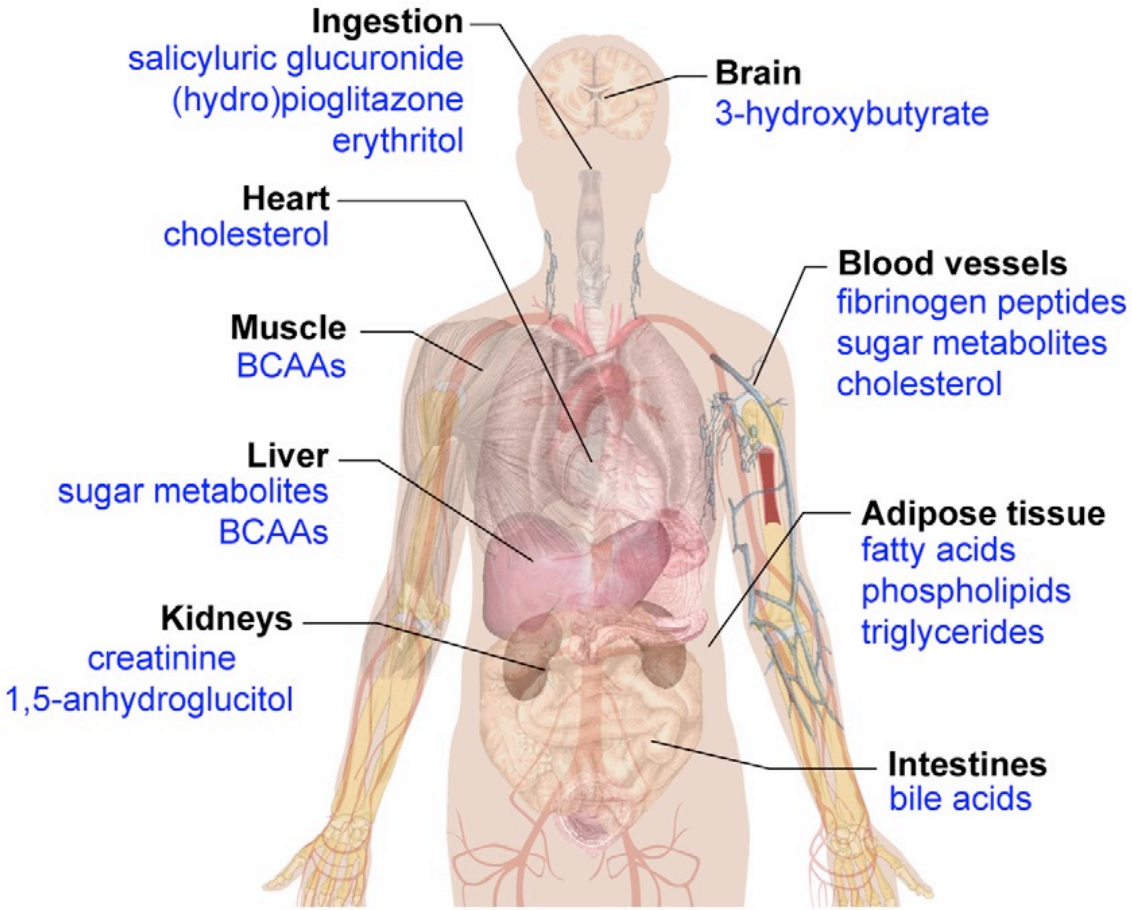
- Exposure: environment (early, late, diet, lifestyle, chemicals, geography), host (genetic background, age), health change over time (disease, biological ageing process)

- Biomarkers: a substance or biological structure that can be measured in the human body and may influence, explain or predict the incidence or outcome of disease

The tissue functions that 1H NMR metabolites represent



Type II Diabetes prediction By 1H NMR metabolites



Suhre et al. PLoS ONE |November 2010 | Volume 5 | Issue 11

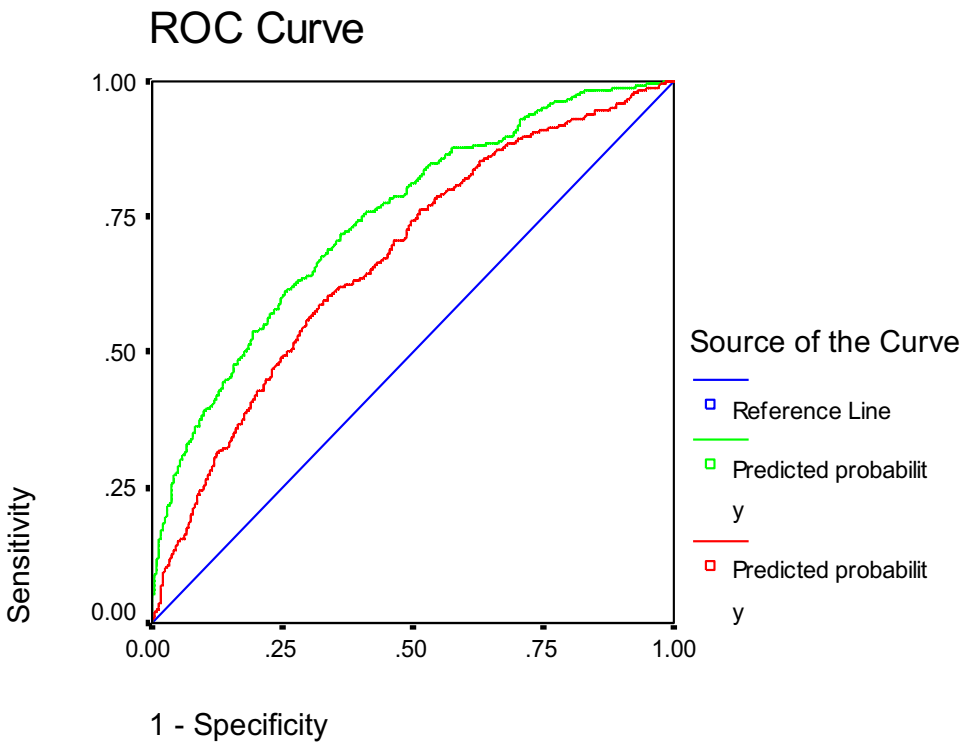
Metabolomics in Research projects

Prediction

Generate a predictor: factor identification and model development :

1. **Exploration of associations between metabolites and diverse endpoints.**
2. **Cross sectional → Prospective/longitudinal follow-up studies**
3. **Univariate (single metabolites) , multivariate**
4. **Replication in independent studies**
5. **Meta-analysis in multiple studies, create predictors (for example of mortality risk) and compare to existing predictors**

Receiver Operator Characteristic (ROC) curves to compare novel and traditional predictors (example 10 y MI risk)



Blue: 50%-50%

→ AUC=0.5

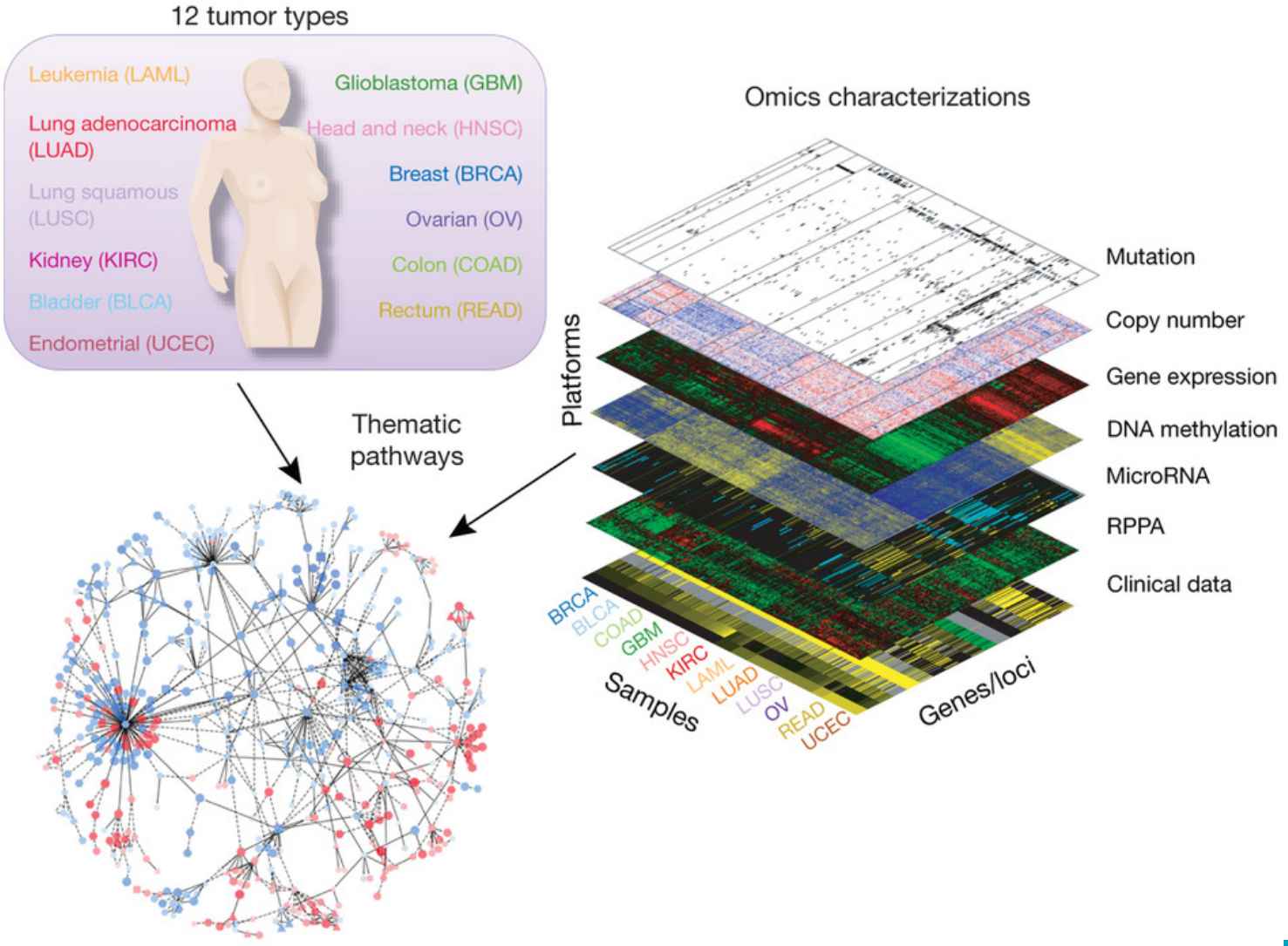
Red: model with Age , Diabetes, Smoking

→ AUC=0.67

Green: model with Age, Diabetes, Smoking HDLcholesterol and systolic blood pressure

→ AUC=0.75

Large scale omics data ; combined to classify patients and predict disease.




shutterstock.com • 790740838